Large Scale Management of Physicists Personal Analysis Data without Employing User and Group Quotas

A. Norman[1], M. Diesbug[1], M. Gheith[1], R. Illingworth[1], A. Lyon[1], M. Mengel[1]

[1]Fermi National Accelerator Laboratory, Bativa IL, USA

E-mail: anorman@fnal.gov

Abstract. The ability of modern HEP experiments to acquire and process unprecedented amounts of data and simulation have lead to an explosion in the volume of information that individual scientists deal with on a daily basis. Explosion has resulted in a need for individuals to generate and keep large personal analysis data sets which represent the skimmed portions of official data collections, pertaining to their specific analysis. While a significant reduction in size compared to the original data, these personal analysis and simulation sets can be many terabytes or 10s of TB in size and consist of 10s of thousands of files. When this personal data is aggregated across the many physicists in a single analysis group or experiment it can represent data volumes on par or exceeding the official production samples which require special data handling techniques to deal with effectively.

In this paper we explore the changes to the Fermilab computing infrastructure and computing models which have been developed to allow experimenters to effectively manage their personal analysis data and other data that falls outside of the typically centrally managed production chains. In particular we describe the models and tools that are being used to provide the modern neutrino experiments like NOvA with storage resources that are sufficient to meet their analysis needs, without imposing specific quotas on users or groups of users. We discuss the storage mechanisms and the caching algorithms that are being used as well as the toolkits are have been developed to allow the users to easily operate with terascale+ datasets.

1. Overview

Fermilab neutrino and muons programs have undergone a substantial ramp up in their computing efforts as part of their analysis, simulation and commissioning efforts. They are no routinely engaging in large scale "production" level data processing, simulation and analysis of data sets and these efforts are generating large amounts of data both in terms of raw byte counts and in terms of total file counts. By way of example, the NOνA In the first 12 months of physics operation, accumulated over 1.6 PB of production level data and simulation which was spread over more than 12 million files. The ramp up for the NOνA experiment shown in FIG. 1 represent only those files which were produced through the physics data taking efforts and the official production and simulation groups.

The storage and management of these data and file volumes is handled through three primary "storage domains" corresponding to conventional random access storage POSIX compliant file system overlays (i.e. systems presenting themselves as "big disks"), high performance storage

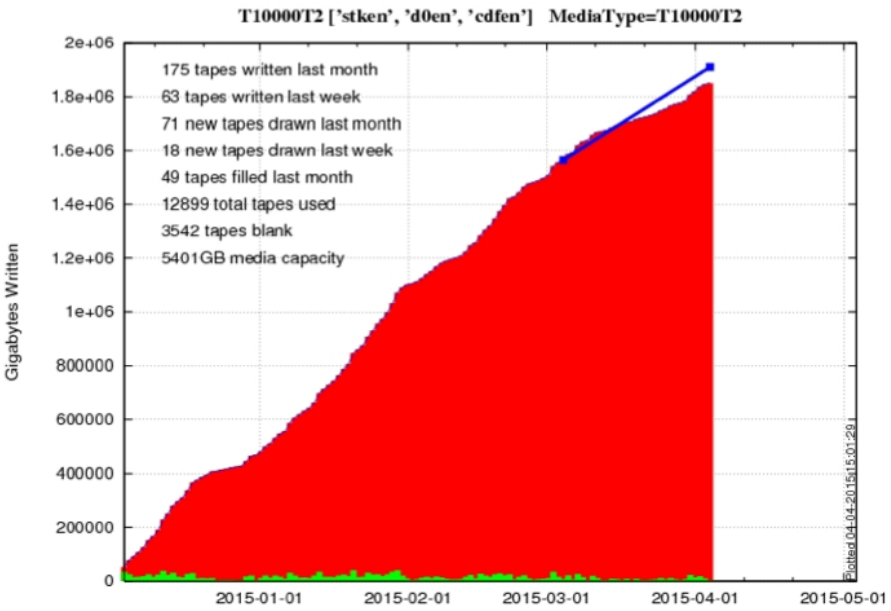

Figure 1. Data volume as a function of time produced and stored by the NOνA experiment

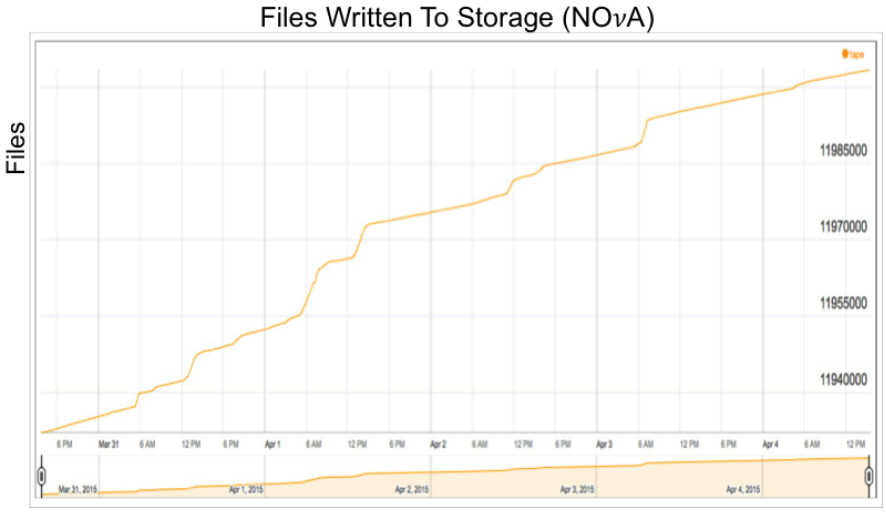

Figure 2. File count as a function of time produced by the NOνA experiment

elements and cache systems, and sequential access archival storage systems (i.e. tape). Each domain is characterized by a set of properties, shown in FIG. 3, which significantly differentiate it from the other domains and which define the tasks for which it is most appropriate.

The problems that arises as experiments scale up their analysis activities are that the data and file volumes associated with the end user analysis efforts quickly begin to rival and eventually exceed the volumes that are stored and managed through the "official" production channels. The access to this data along with its organization, management and scalability are issues which need to be addressed through an evolution of the storage models and tools.

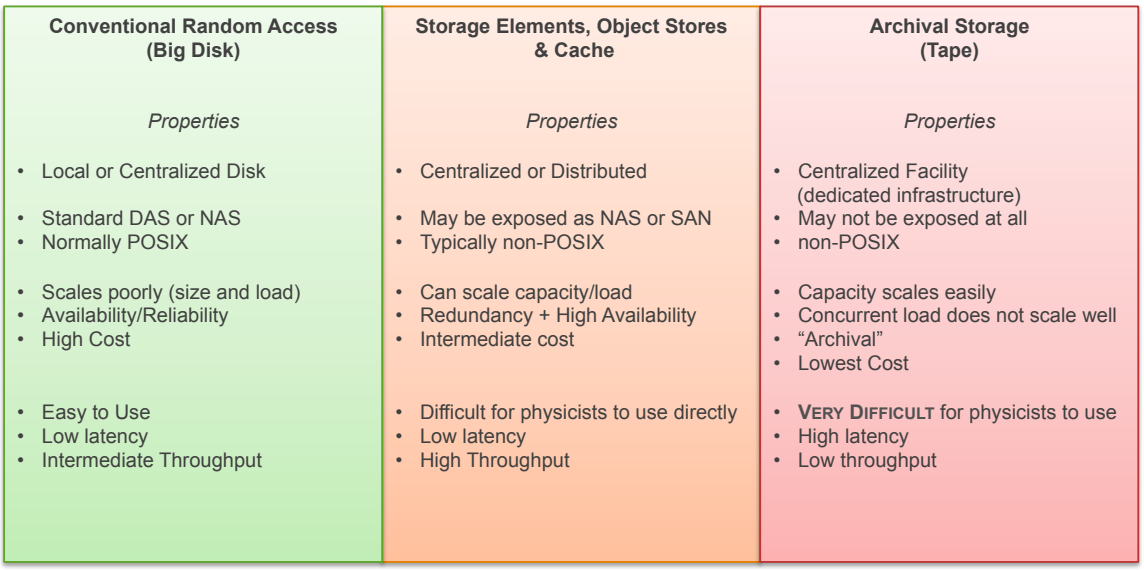

Figure 3. Generalized storage domains that are used in the storage, analysis and archiving of HEP data. Each domain can be characterized and differentiated from the other domains by a number of properties specific to its underlying technologies

2. Storage Domains

The typical physicist performing analysis at Fermilab has access to each of the three domains described in FIG. 3 through a set of centrally managed services. At Fermilab the large conventional random access (POSIX) storage services are provided through a set of central NAS's based on BlueArc Titan servers. These NAS servers provide over 1.2 PB of storage to the experiments, presented as standard NFS exported volumes which are mounted by the experiments both in their interactive analysis environments, as well as across the Fermilab batch farms. Each of the servers provides an network pathway to other Fermilab computing resources with a maximum aggregated bandwidth of 6/4 Gbps. The system is extremely easy for analysis users read and write data to, but experiences performance degradation under high concurrent loads, like those seen when large clusters of batch analysis jobs are run either reading from or copying output back to the NAS.

The majority of analysis users want to interact with their data in this storage domain and for some types of work, typically working with small analysis ntuples and data summary files, the NAS systems provide easy integration with existing analysis tools, low latency access and moderate data throughput.

In contrast to the BlueArc NAS system, a set of large high performance centralized storage elements are provided through dCache based and aggregated disk pools. The dCache systems at Fermilab provide over 4 PB of cache, durable and volatile storage to the physics community. The dCache systems present a unified file system to the users but do not provide POSIX compliant access to the storage. In particular the dCache systems used by the experiments are accessible through the gridftp protocol, an SRM interface, dccp and the xrootd protocol but the files are not directly readable or modifiable through the file system layer. The files stored in dCache are also immutable (non-modifiable) after the initial write into the system, causing the user to copy the file out of the primary storage and to the local disk of their compute node for analysis. The advantage of the dCache storage systems are that they can handled high concurrent loads both for read and write access, have higher aggregated network bandwidth available to them than the NAS systems and support the use of cache policies to mange the disk volumes.

Table 1. NOνA dataset characteristics for first analysis work.

Type	Spills	Files	Size/TB
Official 1^{st} Analysis Dataset	14,308,325	166,629	6.51
User Studies, Merges	study dependent	5-10k	6.5-10
User Skims/NTuples	analysis dependent	5-10k	1-2

The majority of analysis users need to interact with their data in this storage domain when they are doing large scale processing in order to obtain the high levels of concurrent disk and network performance that are required to deliver the data at a fast enough rate to keep the processing efficiencies, as measured by the ratio of CPU time to wall clock time, of analysis jobs high.

The dCache systems also serve as a front end for the archival storage system which is an enstore tape library. The tape library is not directly exposed to the analysis users. Instead the archival storage is exposed only through the data catalogs and data handling systems that the experiments use. This prevents accidental interactions with the high latency tape robots which can be detrimental to the overall performance of the system, while simultaneously allowing for optimizations to be performed when doing writes and restores to the tapes.

The majority of final data sets need to be stored into the tape systems due to their need for long term storage and an archival medium and the high storage capacity and low cost that the tape systems offer.

The current collection of Fermilab developed and supported tools that are part of the FIFE[1] suite provide much of the infrastructure and interface layers that allow for analysis projects to use all of these storage domains successfully. In particular the combination of the SAM[2] data management system[1], the IFDH[3][2] and the Fermi File Transfer Service (F-FTS) work together and have been integrated with the art analysis framework and with the job submission system to the Fermilab and Open Science Grid batch systems to fully exploit the analysis of data which is homed on the different Fermilab storage systems. This has allowed for the successful migration of the large scale production processing for experiments likeNOνA, Minerva away from the central NAS volumes to operate instead from the high speed storage dCache storage elements with permanent storage to the archival tape systems.

The data that uses these data management systems is fully described in the data catalogs making it easy to audit the different types of data, their parentage, sizes and locations as well as many other characteristics which are needed to manage peta-scale volumes of information.

3. User Analysis Data

The current challenge of large scale data management for many of the running experiments at Fermilab, is how to effectively provide the same level of management that is available to the official production data samples, to the data being generated by the individual users doing analysis.

One example of this challenge can be seen through the examination of the datasets being analyzed for the NOνA experiment's first analysis of far detector beam data. The NOνA experiments official data set represents approximately 7 months running with the full far detector[3]. The signal data, representing the time windows of detector readout that overlap

[1] Fabric for Frontier Experiments
[2] Sequential Access with Metadata
[3] Intensity Frontier Data Handling

with the NuMI accelerator beam spills, after full event reconstruction and particle identification processing, constitutes 6.51 TB of data spread over 166k files, as shown in Table 1. This represents only potential signal, while the fully reconstructed data sample used for background determination, calibration and tuning is ten times the size of the beam data[4]. These two samples, the signal and background, are the starting point for most analysis work.

The typical analysis chain starts with the official data samples and from these are derived additional samples for specific studies. This process often involves merging the data down into fewer files but at the same time frequently involves as much as doubling the size of the events through the extra information that is generated for the studies. This typically causes the derived data to be the same size or greater than the original official datasets, but spread over fewer files. The next step in the analysis chain then involves skimming and slimming the information to select out only events and data products of interest and often a conversion of the data into standard Ntuple formats which are easier to work with for a given study or analysis. This results in a significant reduction in the size of the data that is written out, typically 1-2 TB, spread over a similar number of files, 5-10k. The end result is that for any given study each physicist generates on the order of 7-12 TB of new data spread over 10-20k files.

These levels of data are not by themselves daunting, but there is also not a single physicist looking at data on each experiment. An experiment like NOνA has over 100 physicists, postdocs and students who have active analysis areas on the central project disks and dCache volatile storage volumes at Fermilab. Auditing of these areas show that over half of these users already have private dedicated skims and ntuples which they have generated. This means that based on our previous estimates when aggregated over the active analysis users on the experiment we should expect to see 0.3-0.6 PB of studies, skims and ntuples spread over 1-2 million files.

This is exactly what we see after auditing the NOνA analysis areas 6 months into the first analysis push.

4. User Data Management

The question arises, "How do you manage all this data?". The answer is that you can not. There are simply too many files in too many locations and there is little to no record of where the files are, what they contain, their parentage and other provenance.

Imposing quotas on the allocation of storage space doesn't work. Quotas are able to limit the amount of information that a person or group of people can store, but they do not impose any organization on the information. When quotas are reached they do not provide any type of auditing or "cleanup" mechanism on the storage, but instead require humans to either adjust the limits or manually manage the space that is being used. These types of cleanup operations are costly in terms of both the man power that is expended during them and the loss of productivity that is experienced during the time between when the quota is reached, and the time when enough space is freed that analysis work can continue. The problem is compounded by the administrative difficulties of needing to evaluate which datasets are more valuable than others and which users should be granted additional resources (at the expense of others).

Instead of relying on quota systems, a system for managing user data was developed. The key requirements that were foremost in its design were that it must be 1) trivial for the end user to use, 2) that it had to integrate seamlessly with existing analysis tools, and 3) that it must allow for automated cleanup and archiving of data. The model that we designed and adopted for this system was a "Data Catalog Lite" design, where each file would be registered and tagged in a full featured data and replica catalog, but would have a simplified record structure that would not mandate that full provenance tracking and metadata information be provided for each file. The

[4] The raw background and calibration samples an additional factor of 10 times larger than the beam data but was not reconstructed for the first analysis results.

catalog would then be fully integrated with the standard analysis tools and framework through the existing data handling modules and http based APIs which the SAM data handling system already uses. To ease the use of the catalog we also designed a set of tools which were "task" specific and encapsulated all the steps and storage domain specific functions that are required to operate against the archival, cache, distributed and traditional storage systems. The tools remove the need to deal with individual files from the common tasks that the physicists perform when doing analysis and when working with their data, and replaces it with instead working with ensembles of files, or "datasets", as a whole.

In implementing this system we started by using the full featured SAM system which has been heavily used by Fermilab experiments since Tevatron Run II. This system is already designed for optimizing file delivery from archival and cache storage systems and is easily extensible to other storage systems. The SAM system had mainly been a large scale "production" tool due to its older architecture requirement, but it has recently been modernized to allow all of the API calls to function over the http protocol and for many of the older architecture restrictions to be removed[4].

For the purposes of working with user level analysis data we further relaxed the requirements on the SAM system and SAMWeb interfaces to provide an easier interface that analysis users could directly use. As part of this relaxation of requirements we removed the strict requirements on user supplied metadata at the time of file registration, and we added additional authentication protections to the interface to prevent users from accidentally interfering with each other's data.

Central to the way that SAM operates is the concept of a dataset, which is simply a collection of files that are logically associated or belong together based on their individual metadata. In large production operations this is done through complicated relational queries based on the physics parameters and provenance data that are associated with each file. For the user version of these tools we relaxed the association process to allow for files of any type to be associated together simply by the user specifying "They belong together because I say they do". While this association by user fiat may seem arbitrary, it provides an extremely powerful tool for organizing both similar and dissimilar data. Users can in effect declare large groups of files to the data catalog and then quickly select subsets of them to operate on, or associate more information with different subsets to create logical hierarchies and organizational structures which are independent of the actual storage devices.

This organization also allows for quick distributed analysis of the data using a "GetNextFile" paradigm in which individual analysis jobs have no a priori knowledge of which file they will receive, but simply ask for the next available file. This allows any analysis system which can issues the HTTP based API calls to use the full data catalog. The data handling system then is responsible for both optimizing and delivering the file to the jobs, which results in a job being able to interact transparently with the storage infrastructure. This interaction is shown in FIG. 4

5. SAM User Tools

The SAM for users tools set starts with the "Add" a dataset tool. This tool in its simplest invocation associates a group of files together as a named dataset based on no metadata other than the users instruction that they belong together for some reason. The user supplies an arbitrary name for the dataset and either a directory path or a list of files or file locations which contain the files of interest. The tool parses the input file list or searches the specified directory for these files. All files in the directories and optionally any subdirectories, are registered with the SAM data catalog and their current replica information corresponding to the found copy of the file is recorded. Name collisions are prevented within the SAM catalog namespace through an automated renaming process the prepends a UUID string to each of the files that are found. This renaming can optionally be replaced with a user defined set of renaming rules, where any

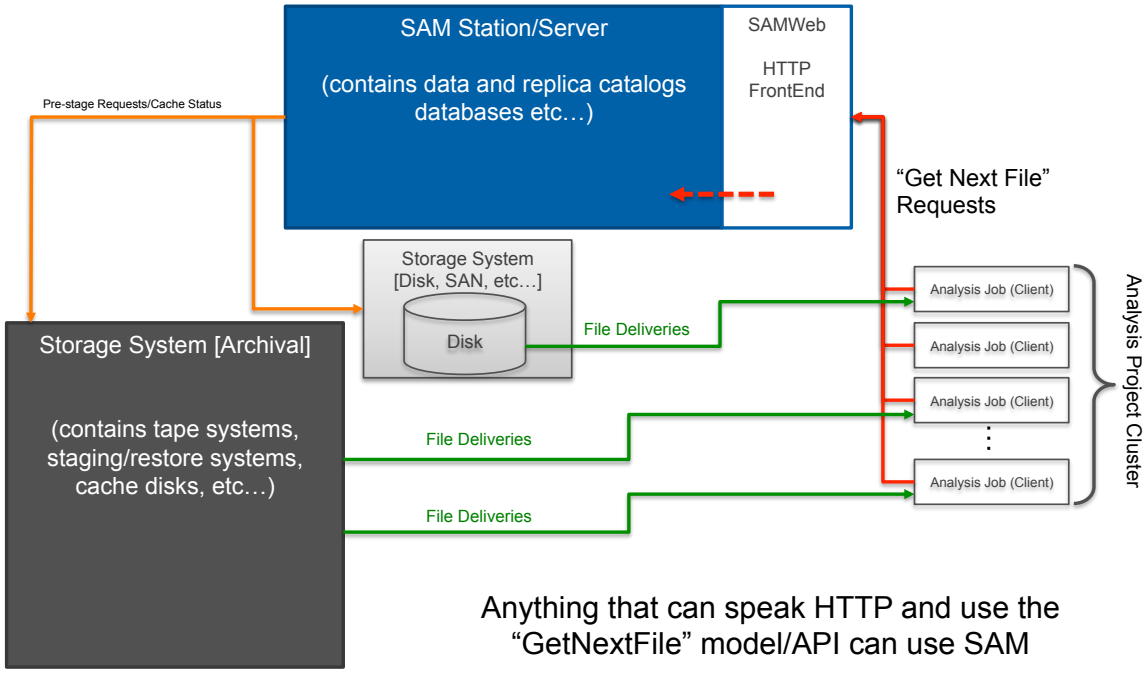

Figure 4. Schematic interaction of distributed analysis jobs with the SAM data handling system and the underlying storage infrastructure.

remaining collisions are left to the user to reconcile. Each of the files are assigned minimal metadata information, corresponding to the dataset tag that was specified by the user, and finally a formal dataset is defined which associates these files through their tag.

The "Add" tool eliminates the confusion of dealing with individual files. The end user never sees or has to deal with the individual registration calls, name collisions or other details of the declaration process. The addition process also scales properly so that the declaration and registration of a dataset with 10's of thousands of files can be processed quickly so that subsequent analysis of the files can being.

The rest of the SAM user tool set consists of the task based functions to perform standard operations on the datasets as a whole. A "clone" and "unclone" tool are provided which create, copy and remove replicas of the dataset on,to and from different managed storage systems. This allows the user to quickly archive their dataset to the tape system, create a low latency copy on the high speed cache disks, or move/remove a copy of their dataset from the NAS systems to more appropriate volumes.

The SAM for user tools also include a "validate" command which audits a given dataset to ensure that all the files are present and accessible. This is used primarily to audit different cache systems to ensure that the dataset in question is fully in the cache prior to starting an analysis pass on the data.

If the user decided that their data should have additional meta information attached to it they are able to quickly associate different metadata parameters to each of the files in the dataset through the "modify_metadata" command. This serves as a bridge for them to use the more sophisticated aspects of the SAM query engines for doing advanced searches and associations of the data.

Finally the "retire" tool provides the functionality to remove files from both the storage and from the data catalog. The tools provides the full functionality to allow the user to cleanup even large datasets quickly and without having to know the specifics of all of the file replicas or the

semantics of how to interact with the storage they are housed on.

6. Conclusions

The SAM for users tool suite removes the need to rely on quota based management for the storage of user data. It removes this need by providing an overall reduction in the complexity of the management process. Instead of the physicists needing to manage millions of individual files, they need to deal only with a handful of dataset names.

This also provides a boost to the operational capabilities that are available to the users by insulating the physicists from having to known and understand the details of how the complicated modern storage systems operate. Instead all they need to only know their dataset's name to analyze it. They are also able to share their dataset with others within the collaboration without those individuals needing to know about the actual file locations or storage details. This expedites the analysis processing and removes many of the problems with knowing exactly which files were used in a given study or analysis.

Finally the SAM for users tools provides automated cleanup functions. The system allows for automated data movement between storage domains, the creation of replicas at different locations, the archiving of data to the tape libraries and ultimately the removal of files from working areas when the users are done with them. It accomplishes this without the physicists needing to know about actual locations of the files or the underling storage that is being used.

The system also works all the different scales which are needed for doing the end to end processing and analysis of modern data. It works for standard analysis jobs which are processed through the official framework and scale into the 100k's of files and many hundreds of terabytes of data. Similarly it works for specialized analysis studies and skims which may user custom analysis frameworks while needing to access 10k's of files or a few terabytes of data. Finally it works even at the final analysis level where the physicists are examining ntuple style data in an interactive ROOT sessions, where chained together trees can be specified and streamed via the xrootd protocol from remote storage elements.

The adoption of the system is already wide spread at Fermilab. In the first week after initial release, over 260 thousand user files were registered with the system and analysis projects with this data had been run.

Acknowledgements

The author acknowledges support for this research was carried out by the Fermilab scientific and technical staff. Fermilab is Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy

References

[1] Lyon A, Illingworth R, Mengel M and Norman A 2012 *J. Phys.: Conf. Ser.* **396** 032069
[2] Lyon A L and Mengel M W 2014 *Journal of Physics: Conference Series* **513** 032068
[3] Norman A 2015 *AIP Conf. Proc.* (Neutrino 2014)
[4] Illingworth R A 2014 *Journal of Physics: Conference Series* **513** 032045